

WHAT IS CLAIMED IS:

1. A method of processing a body of text to generate compression options, comprising:
 - performing a linguistic analysis on the body of
5 text to obtain a linguistic output
indicative of linguistic components of the
body of text; and
 - generating a plurality of compression options to
compress the body of text based on the
10 linguistic output.
2. The method of claim 1 wherein generating a plurality of compression options comprises:
 - 15 subjecting a portion of the body of text to
different sets of compression rules to
obtain the plurality of compression
options.
3. The method of claim 2 wherein subjecting the
20 body of text to different sets of compression rules,
comprises:
 - subjecting the portion of the body of text to
the different sets of compression rules in
a predetermined order such that the
25 compression options reflect varying degrees
of compression of a same portion of the
body of text.
4. The method of claim 4 wherein generating a
30 plurality of compression options comprises:

generating a compression identifier attribute
indicative of at least one of the sets of
compression rules to which the portion of
the body of text is subjected.

5

5. The method of claim 4 wherein generating a
plurality of compression options comprises:

generating a ShortForm attribute indicative of a
compressed form of the portion of the body
10 of text after application of the set of
compression rules.

6. The method of claim 5 wherein generating a
plurality of compression options comprises:

15 generating a case normalized attribute, based on
the ShortForm attribute, indicative of a
CaseNormalizedForm of the ShortForm
attribute.

20 7. The method of claim 6 wherein generating a
plurality of compression options comprises:

generating a compression attribute indicative of
a further compressed form of the case
normalized attribute.

25

8. The method of claim 7 wherein generating a
compression attribute comprises:

applying letter removal rules to the case
normalized attribute to remove letters

based on a predetermined location of the
letters in the CaseNormalizedForm.

9. The method of claim 8 wherein generating a
5 plurality of compression options comprises:
generating a LongForm attribute that reflects
substantially no compression of the portion
of the body of text.

10 10. The method of claim 9 wherein one ShortForm
attribute comprises a word substitution based on a
dictionary look-up and wherein generating a plurality
of compression options comprises:
setting the case normalized attribute and the
15 compression attribute to the ShortForm
attribute.

11. The method of claim 5 wherein performing a
linguistic analysis comprises performing a syntactic
20 analysis on the portion of the body of text and
wherein generating the ShortForm attribute comprises:
applying the set of compression rules based
on the syntactic analysis.

25 12. The method of claim 11 wherein the linguistic
analysis further comprises, prior to performing the
syntactic analysis:
performing a lexical analysis on the body of
text; and

performing a morphological analysis on the body
of text.

13. The method of claim 5 wherein generating the
5 ShortForm attribute comprises:
normalizing dates to a numerical form.

14. The method of claim 5 wherein generating the
ShortForm attribute comprises:
10 normalizing offset dates to a numerical form,
based on a date that the body of text was
authored.

15. The method of claim 5 wherein generating the
15 ShortForm attribute comprises:
maintaining symbol-sensitive text fragments in
uncompressed form.

16. The method of claim 15 wherein maintaining
20 symbol-sensitive text fragments comprises:
maintaining text fragments that, cannot be
accurately understood unless maintained
fully in-tact, in uncompressed form.

25 17. The method of claim 16 wherein maintaining text
fragments comprises:
maintaining uniform resource locators and
electronic mail addresses in uncompressed
form.

18. The method of claim 11 wherein the syntactic analysis includes a tree having non-terminal nodes representing multi-word portions of the body of text and terminal nodes indicative of words in the body of text, and wherein both the non-terminal nodes and the terminal nodes are examined for application of compression rules.

19. A data structure formed from an analysis of a portion of a body of text indicative of a plurality of compressed forms of the portion of the body of text, the data structure comprising:

a plurality of data fields, representing a plurality of compressed forms of the portion of the body of text.

20. The data structure of claim 19 and further comprising:

a compression type attribute indicative of a type of compression applied to the portion of the body of text in generating at least one of the plurality of compressed forms.

21. The data structure of claim 20 wherein the plurality of compressed forms comprises:

a ShortForm attribute indicative of a compressed form of the portion of the body of text after application of the type of compression identified by the compression type attribute.

22. The data structure of claim 21 wherein the plurality of compressed forms comprises:

5 a case normalized attribute, based on the ShortForm attribute, indicative of a CaseNormalizedForm of the ShortForm attribute.

10 23. The data structure of claim 22 wherein the plurality of compressed forms comprises:

a compression attribute indicative of a further compressed form of the case normalized attribute.

15 24. The data structure of claim 23 and further comprising:

a LongForm attribute indicative of substantially no compression of the portion of the body of text.

20

25. A message handler receiving a message and generating compression options indicative of different forms a portion of a body of text in the message, the message handler comprising:

25 a linguistic analyzer linguistically configured to analyze the body of text and provide a linguistic analysis; and
a compression form generator configured to generate a plurality of compressed forms of

a portion of the body of text based on the linguistic analysis.

26. The message handler of claim 25 wherein the
5 compression form generator is configured to apply a plurality of different sets of compression rules to the portion of the body of text obtain the plurality of compressed forms.

10 27. The message handler of claim 26 wherein the compression form generator is further configured to apply the different sets of compression rules in a predetermined order such that the plurality of compressed forms reflect varying degrees of
15 compression of a same portion of the body of text.

28. The message handler of claim 27 wherein the compression form generator is further configured to generate a compression identifier attribute
20 indicative of at least one of the sets of compression rules applied to the portion of the body of text.

29. The message handler of claim 27 wherein the compression form generator is configured to provide,
25 at its output, a data structure containing a plurality of attributes indicative of the plurality of compressed forms, and the compression identifier attribute.

30. The message handler of claim 29 wherein the plurality of attributes includes:

5 a ShortForm attribute indicative of a compressed form of the portion of the body of text after application of the set of compression rules;

10 a case normalized attribute, based on the ShortForm attribute, indicative of a CaseNormalizedForm of the ShortForm attribute; and

a compression attribute indicative of a further compressed form of the case normalized attribute.

15 31. The message handler of claim 30 wherein the plurality of attributes further comprises:

a LongForm attribute that reflects substantially no compression of the portion of the body of text.